GEOG0013 Data Analysis Project Worksheet 2022/23

Workshop 1: Introduction to Statistics

Before you begin

To complete the worksheet, you will need to make use of the datasets, techniques, and R code you have worked with on the self-guided RStudio practicals in Week 4 which are hosted **[here]**. The worksheet can be completed as you work through the self-guided practicals OR you can go through the guidance notes in the practicals first and then complete the worksheet – either way it is up to you.

Background

You are tasked with conducting an overall descriptive analysis of various socioeconomic indicators across 73 Spanish neighbourhoods in Barcelona using the secondary datasets provided for the Week 4's workshop exercise. These are available for download **[here]**.

You will have access to a set of files:

1. **Barcelona_lifeexp_2015.csv** which contains information on averaged life expectancy across neighbourhoods from 2010-14

2. **Barcelona_unemp_2015.csv** which contains information on unemployment [%] across neighbourhoods in 2015

3. **Barcelona_fcitizens_2015.csv** which contains information on a neighbourhood's population who are foreign nationals [%] in 2015

4. **Barcelona_income_2015.csv** which contains information on the average neighbourhood family income relative to Barcelona average (2015). Values of 100 indicate Barcelona's average income, <100 denotes below average incomes, values >100 denote above average incomes.

5. **Barcelona_list_of_districts.csv** which contains the list of all districts (as well as their neighbourhoods) in Barcelona.

Tasks

1. Pick one variable to focus on from those available in the csv files you have downloaded that you are interested in analysing. Do not use the rents and cars as this dataset was used in the video tutorials!

2. Using appropriate techniques, describe its frequency distribution, and overall summary measures of your selected variable. Hint: Use the techniques taught in the guide at your disposal to complete this task.

3. Select **two** districts that have at least 7 neighbourhoods and perform a comparative analysis using summary measures for your chosen variable. Provide the following: 1.) 3 to 5 lines of code chunks showing this data cleaning was performed; 2.) an interpretation of the result showing the differences.

4. Use the techniques at your disposal to make an overall descriptive comparison between low- and high-income neighbourhoods using your variable of interest. Are there any differences?

Format and submission

You are to work individually to complete these tasks. Your answers must be written in a minimum of font size 11 and your submission document must be no longer than 2 sides of A4 (including everything). The usual departmental penalties will be applied to overlength work. Don't forget to provide informative titles to your figures, with appropriately labelling and size them suitably. For tabular results, you are expected to produce them in a formatted style. Do not copy and paste an R output and leave the result unformatted (you will lose marks if results are presented in an unacceptable format!).

The submission deadline is noon on Monday 5th December 2022. Submit the worksheet through Turnitin on the module Moodle page. Please make sure to enter your student exam candidate ID at the top of the page and use it as your submission title on Moodle. This worksheet counts for 10% of your GEOG0013 grade and will be marked using the same grading matrix used for the other worksheet tasks. Remember that 60% of your mark comes from the Field Notebook. Please see the Moodle guidance page (in the Week 4 section of the GEOG0013 page) for details of what to include in the notebook from the Data Analysis project.